Prekladateľská činnosť

Radev, M.: Na ulici. Bratislava: Mladé letá 1962. 98 [1] s.

Čašule, K.: Hra na manželstvo. Veselohra zo súčasného života v troch dejstvách (siedmich obrazoch). Bratislava: Diliza 1967. 52 [1] s.

Vukasovič, A.: Diferenciácia systému pedagogických vied (príspevok do diskusie). In: Pedagogická revue, 1995, roč. 47, č. 1 – 2, s. 8 – 16.

Čukan, J. – Vančo, J. – Chlebnický, J.: Dolnozemské reflexie na neroľnícke zamestnania. Nitra: Univerzita Konštantína Filozofa 2001. 231 s. ISBN 80-8050-422-9.

Redakčná činnosť

Slavica Slovaca, 1988, roč. 23 – 1998, roč. 33 (člen red. rady).

Slavica Slovaca, 1999, roč. 34 – 2002, roč. 37 (hlavný redaktor).

Slavica Slovaca, 2005, roč. 40 – 2007, roč. 42 (člen red. rady).

Prameň, tlačový orgán Matice slovenskej v Chorvátsku (lektor)

Karadžic, V. S.: Ti prstene. Bratislava: Tatran 1988. 240 s. ISBN 86-7103-033-4 (zostavovateľ).

Zborník Spolku vojvodinských slovakistov. 14 (1992). Nový Sad (Juhoslávia), Spolok vojvodinských slovakistov 1997. 276 s. (redaktor).

Slovensko-chorvátske jazykové a literárne vzťahy. Zborník prác z medzinárodnej vedeckej konferencie Slovensko-chorvátske jazykové a literárne vzťahy, ktorá sa uskutočnila 22. – 23. apríla 1999 v Bratislave. Bratislava, T. R. I. Médium pre Združenie slovanskej vzájomnosti v Bratislave 1999. 160 s. ISBN 80-88676-18-5 (zostavovateľ).

Zostavili: *Ladislav Dvonč, Júlia Behýlová*

SLOVKO 2007
POČÍTAČOVÉ SPRACOVANIE PRIRODZENÉHO JAZYKA, POČÍTAČOVÁ LEXIKOGRAFIA A TERMINOLÓGIA

V dňoch 25. – 27. októbra 2007 sa v priestoroch Inštitútu pre verejnú správu na ul. M. Schneidra-Trnavského 1 v Bratislave uskutočnil v poradí štvrtý medzinárodný seminár SLOVKO 2007, ktorý organizovalo oddelenie Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied v Bratislave. Témou tohtoročného seminára bolo *Počítačové spracovanie prirodzeného jazyka, počítačová lexikografia a terminológia*.

Vysokú odbornú úroveň garantovali 35 prednášatelia prevažne zo strednej a východnej Európy (Slovensko, Česko, Poľsko, Ukrajina, Rusko, Chorvátsko), ale aj účastníci zo Španielska či Belgicka. Vzhľadom na to, že išlo o seminár medzinárodného charakteru, prednášky aj diskusné príspevky odzneli zväčša v angličtine, ale aj v slovenčine, češtine, ruštine a ukrajinčine. (Väčšina z nich je publikovaná v zborníku *Computer Treatment of Slavic and*

East European Languages, Fourth International Seminar. Eds. J. Levická – R. Garabík. Brno: Tribun 2007. 318 s. ISBN 978-80-87139-05-9). Na konferencii bolo prítomných viac ako 50 účastníkov a odznelo 35 príspevkov.

Po otvorení konferencie Janou L e v i c k o u, pracovníčkou Slovenského národného korpusu JÚĽŠ SAV v Bratislave, ako prvý vystúpil Luis V i l l a r e j o z Open University of Catalonia s príspevkom *Lexterm, an Open Source Tool for Lexical Extraction.* Hovoril o vývoji a využití nástroja na vyhľadávanie termínov a terminologických jednotiek zvaného Lexterm, ktorý vypracovali štyri katalánske univerzity v rámci projektu RESTAD. Cieľom tohto projektu je zautomatizovať proces prekladania akademických dokumentov z katalánčiny do španielčiny a angličtiny, poskytnúť prekladové ekvivalenty z dvojjazyčných korpusov a vytvoriť tak predpoklady pre automatický alebo poloautomatický preklad textov.

Dominika U r b a ń s k a z Poľsko-japonského inštitútu informačnej technológie sa zaoberala témou *Automatic Term Recognition in Polish Texts*. V súvislosti s automatickým vyhľadávaním termínov v databáze uviedla, že nakoľko je poľština jazyk s bohatou flexiou, lingvistický ani štatistický prístup neprinášajú uspokojivé výsledky. Pri práci používajú zmiešanú metódu pozostávajúcu z gramatického filtra a zo štatistického prístupu založeného na modifikovanom Cohenovom algoritme. Ich cieľom je vytvoriť efektívny nástroj na pomoc ďalšiemu lingvistickému výskumu.

Ako tretia vystúpila Oksana S. K o z a k z Kyjevskej národnej lingvistickej univerzity s príspevkom *The Role of Word Frequency Vocabularies in the Research of Psychology and Philosophy Terminological Systems*. Poukázala na spôsob využitia frekvenčných slovníkov pri analýze termínov z oblasti filozofie a psychológie, pričom vychádzala z výskumu odborných časopisov z konca 19. a začiatku 21. storočia.

Ďalšou témou bola prednáška Michala K ř e n a z ÚČNK Karlovej univerzity v Prahe *Variation of Czech Lexicon as Reflected by Corpora Comparison.* Autor predstavil Český národný korpus, reprezentatívne vyváženú elektronickú databázu obsahujúcu 100 miliónov textových slov. Keďže korpus zachytáva dve na seba nadväzujúce časové obdobia: SYN 2000 (od roku 1990) a SYN 2005 (od roku 2000), cieľom výskumu je frekvenčné porovnávanie tokenov v oboch korpusoch, ktoré má stanoviť, či v používaní českých slov obsiahnutých v týchto dvoch korpusoch nastali signifikantné rozdiely.

Marek G r á c z Fakulty informatiky Masarykovej univerzity v Brne sa vo svojom príspevku *Effective Methods of Building Slovak-Czech Dictionary* venoval problematike strojového prekladu medzi češtinou a slovenčinou. Poukázal na to, že existujúce slovníky zachytávajú predovšetkým dvojice slov, ktoré sa navzájom líšia, a navrhuje rozšíriť ich o ľahko zameniteľné dvojice, pričom podstatou navrhovanej metódy je nájsť vhodné prekladové ekvivalenty pre dané slovo a vybrať najvhodnejšiu lemu z druhého jazyka ako preklad. Predbežné výsledky potvrdzujú, že tento prístup zlepšuje efektívnosť pri vytváraní česko-slovenského slovníka.

Aleš H o r á k a Adam R a m b o u s e k z Fakulty informatiky Masarykovej univerzity v Brne sa prezentovali prednáškou *Administration Framework for the DEB Dictionary Server.*

V príspevku predstavili implementáciu nástrojov na správu systému DEB II slúžiaceho na písanie slovníkov. Opisovali používanie nástroja na správu užívateľov a názorne ukázali, ako pomocou týchto nástrojov systém nastavovať a adaptovať na nové slovníky. Informovali tiež o svojej práci na automatickom generovaní užívateľských rozhraní na základe ich zovšeobecneného popisu.

Ako ďalší vystúpil Marek C i g l a n z Ústavu informatiky SAV v Bratislave a predniesol za kolektív autorov prednášku: *Semi-automatic Semantic Annotation of Slovak Texts*, v ktorej uviedol, že doterajšie metódy sémantickej anotácie textov sa dali aplikovať najmä v angličtine a neboli vhodné na vysoko flektívne jazyky ako slovenčina. Ich kolektív preto vyvinul špecifický nástroj Ontea vhodný na poloautomatickú sémantickú anotáciu slovenských textov založenú na skúmaní bežných výrazov, ktorý spolu s nástrojmi na prirodzenú jazykovú identifikáciu, lematizáciu a stematizáciu a pomocou špeciálneho indexového mechanizmu poskytuje sľubné výsledky pre sémantickú anotáciu slovenských textov. Prehodnotením tejto metódy sa dosiahla 70-percentná úspešnosť.

Dana H l a v á č k o v á a Karel P a l a z Fakulty informatiky Masarykovej univerzity v Brne v prednáške *Computer Processing Derivational Relations in* Czech predstavili prvé výsledky komputačnej analýzy základných a najproduktívnejších derivačných vzťahov v češtine, ku ktorým dospeli pomocou derivačného webového rozhrania derivačnej verzie morfologického analyzátora Ajka. Autori vyvinuli špeciálne derivačné rozhranie, prostredníctvom ktorého skúmali sémantické vlastnosti vybraných menných derivačných sufixov, ako aj slovesných prefixov a vytvorili súbor sémanticky označených derivačných vzťahov (dosiaľ spracovali 14 z 22). V rámci aplikácie pridali vybrané derivačné vzťahy do elektronickej databázy WordNet a tým ju obohatili približne o 30 000 nových českých synsetov.

Dorota V a s i l i š i n o v á a Radovan G a r a b í k zo SNK JÚĽŠ SAV v Bratislave pripravili prednášku *Parallel French-Slovak Corpus,* v ktorej R. Garabík uviedol, že predstavovaný francúzsko-slovenský paralelný korpus FRASK je rozsiahly korpus obsahujúci beletriu a legislatívne texty Európskej únie v obidvoch jazykoch. Texty majú vetnú štruktúru, sú lematizované a obsahujú morfologické informácie. Vyhľadávací mechanizmus zahŕňa možnosť vyhľadávať jednotlivé slová, slovné spojenia, lemy a morfologické údaje používajúc bežné výrazy. Korpus sa naďalej rozširuje a je dostupný verejnosti na internete.

Renáta B l a t n á z ÚČNK Karlovej univerzity v Prahe sa v príspevku *On Valency of Some Czech Verbs with Multi-Word Prepositions (based on the Czech National Corpus*) venovala problematike valencie niektorých českých slovies. Korpusová lingvistika poskytuje nové možnosti skúmania predovšetkým syntagmatických vzťahov v jazyku. Cieľom tejto práce je skúmanie valencie skupiny slovies vyskytujúcich sa predovšetkým v odborných textoch, po ktorých nasledujú predložkové spojenia. Autorka vychádza z lexikologickej koncepcie valencie opísanej F. Čermákom. Jej analýza sa zakladá na výbere 16 najfrekventovanejších predložkových spojení z korpusu českého jazyka SYN2000, pričom skúma ich kolokabilitu vzhľadom na slovesá. Dochádza k záveru, že najčastejšie predložkové spojenia sa vyskytujú v kontexte slovies s intelektuálnym významom.

Radovan G a r a b í k zo SNK JÚĽŠ SAV v Bratislave (v spolupráci s Marínou M i k u l a j o v o u z Pedagogickej fakulty UK v Bratislave) pripravil prednášku na tému: *A Cross-linguistic Database of Children's Printed Words in Three Slavic Languages*. Predstavil lexikálnu databázu pozostávajúcu z morfologicky a foneticky anotovaných slov, ktoré sa najčastejšie vyskytujú v českých, poľských a slovenských učebniciach pre nižší stupeň základných škôl. Korpus je ľahko prístupný cez internetové rozhranie umožňujúce vyhľadávanie bežných výrazov pomocou slov, liem, morfologických značiek a fonemického prepisu. Očakáva sa, že táto databáza bude mať široké využitie aj pri tvorbe experimentálneho materiálu pri psycholingvistickom výskume.

Poslednou prednáškou tohto seminárneho dňa bol príspevok Karla P a l u z Fakulty informatiky Masarykovej univerzity v Brne *Lexicographical Software Tools.* Autor venoval, špeciálne na požiadanie slovenských lexikografiek, hodinovú prednášku lexikografickým nástrojom používaným v českom lexikografickom prostredí. Ako prvý predstavil nástroj Debdict slúžiaci ako browser, pomocou ktorého má lexikograf a iný používateľ prístup ku všetkým existujúcim českým elektronickým slovníkom. Okrem toho v ňom nájdeme odkazy na Encyklopédiu Diderot, Seznam Encyclopediu, Google a Wikipediu, ako aj odkaz na český morfologický analyzér Ajka. Nástroj Derivational web interface (DWI) umožňuje vyhľadávať slová podľa požadovaných kritérií (napr. získavanie zoznamov slov podľa zadaných sufixov alebo prefixov), je v integrácii s Ajkou a obsahuje aj odkaz na Debdict. Nástroj Debterm predstavuje terminologickú databázu špecializovanú na termíny z oblasti umenia. Výnimočnosťou tejto databázy je, že okrem terminologických vstupov obsahuje obrazové dáta a videá a je prístupná v štyroch jazykoch. Nástroj DebVisdic slúži ako browser, ktorý umožňuje prístup k všetkým dostupným Wordnetom. A napokon nástroj Word Sketch Engine, ktorý umožňuje vyhľadávať všetky kontexty a štatistiky požadovaného slova, ako aj jeho gramatické vzťahy s inými slovami, preto sú pre vytvorenie tohto nástroja nevyhnutné morfologicky anotované korpusy.

Prvý blok piatkového seminárneho dňa moderovala Klára Buzássyová. Uviedla Albenu Rangelovú z Ústavu pro jazyk český AV ČR s prednáškou *Wider Framework of the Research Plan Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century.* Autorka informovala účastníkov, že predmetom výskumnej činnosti lexikografickо-terminologického oddelenia Ústavu pro jazyk český v rokoch 2005 –2010 je vytvorenie databázy lexikálnej zásoby začiatku 21. storočia. Zameriava sa na vytvorenie materiálových a technických predpokladov na využívanie moderných informačných technológií pri výskume a popise slovnej zásoby, a v spolupráci s Fakultou informatiky MU v Brne i na návrh a realizáciu vlastného lexikografického softvéru až po sprevádzkovanie navrhnutej lexikálnej databázy na vlastnom serveri.

V bloku českých lexikografov vystúpila aj Jindra S v ě t l á z Ústavu pro jazyk český AV ČR v Prahe s príspevkom *The Possibilities and Limits of Lexicographical Description of the Czech Lexicon in Database Form.* Hovorila o tom, že Ústav pro jazyk český zabezpečuje realizáciu celej lexikálnej databázy po koncepčnej aj obsahovej stránke, pracovníci Fakulty informatiky MU v Brne zabezpečujú podľa požiadaviek pracovníkov Lexikograficko-terminologického oddelenia ÚJČ AV ČR programovú časť, t. j. vytvorenie príslušného softvérového programu, prázdneho jadra databázy a rozhrania pre zápis dát, ktoré by malo po roku 2010 slúžiť aj pre spracovanie nového výkladového slovníka formou lexikálnej databázy.

Po nej vystúpila Milada V o b o r s k á z Ústavu pro jazyk český AV ČR v Prahe s príspevkom *Tools for the Input of Morphological Data – LEXIKON 21 Solution Proposal,* v ktorom sa podrobne zaoberala morfologickými nástrojmi používanými v rámci programu PRALED. Tieto nástroje vyhovujú požiadavkám opisu morfologickej charakteristiky slova, ak berieme do úvahy nejasné hranice medzi slovnými druhmi a ich podkategóriami. Týmto spôsobom bude možné uchopiť a opísať zmeny, variantnosť rodu, nejasnosť čísla pri podstatných menách alebo nájsť slovo s podobnou charakteristikou, ale zaradené do iného slovného druhu.

V bloku českých lexikografov pokračovali Zdeňka O p a v s k á a Barbora Š t ě p á n k o v á z Ústavu pro jazyk český AV ČR v Prahe s prednáškou *Tools for Working with Corpus*

Evidence in the Lexical Database LEXIKON 21(Program PRAMAT and the Exemplification Tool). Lexikograficko-terminologické oddelenie Ústavu pro jazyk český vypracovalo program PRAMAT určený na triedenie materiálov z korpusu, prípadne ďalších textov. Tento program slúži ako „pracovná plocha" na prácu s príkladmi pri vytváraní hesiel v lexikálnej databáze LEXIKON 21. Vzhľadom na komplikovanú štruktúru lexikálnej databázy vznikol špeciálny exemplifikačný nástroj PRALED umožňujúci dôkladnejšiu segmentáciu hesiel ako v bežných jednojazyčných slovníkoch. Táto prednáška sa zameriava na opis a ukážku fungovania týchto nástrojov: výber dokladu z korpusu, pridanie komentára, segmentovanie v PRAMAT-e, vytvorenie a uloženie vybraných príkladov pre budúce spracovanie v LEXIKON-e 21.

Veľmi informatívne dopoludnie pripravené českými lexikografmi uzavrel príspevok Věry C h u d o m e l o v e j a Edith B i r k h a h n o v e j z Ústavu pro jazyk český AV ČR v Prahe:

The Possibilities of Lexicographic Description of Terms in the Lexical Database LEXIKON 21. Uviedli v ňom, že elektronické spracovanie slovnej zásoby každého jazyka poskytuje nové možnosti vrátane systematickejšieho opisu terminologického slovníka, ktorý okrem iného umožňuje presnejší a komplexnejší opis charakteristík termínov, ako je výklad významu, encyklopedická poznámka, podvýznam, exemplifikácia a doplnená informácia v podobe nezávislých nástrojov, medziiným aj zoznam špecializovaných odborov a oblastí.

Sekciu uvádzanú Victorom Zacharovom otvoril František Č e r m á k, riaditeľ Ústavu Českého národného korpusu Karlovej univerzity v Prahe, svojím príspevkom *Úzus a syntax interjekcí: případ češtiny*. V ňom na dátach reprezentatívneho Pražského hovoreného korpusu skúmal aspekty niektorých dosiaľ neskúmaných citosloviec, ich vetnú funkciu a tiež kombinatorickú schopnosť. Povahu jednotlivých distribučných typov posudzoval na pozadí ich sémantickej a funkčnej povahy. Analýzou ich pozičnej a kombinatorickej distribúcie vo vete ukázal, že citoslovce nie sú vo vete izolovanými lexémami vkladanými do propozície bez väzby a súvislosti.

S tematicky príbuzným príspevkom *Partikuly a interjekcie v slovníkových výkladoch* vystúpila vedúca oddelenia SNK JÚĽŠ SAV Mária Š i m k o v á. Autorka krátko predstavila históriu častíc a citosloviec a poukázala na ich postupné vyčleňovanie sa v rámci slovných druhov. Tiež podala stručný prehľad spracovania partikúl a interjekcií v starších slovenských slovníkoch, ako aj ich dnešnej podobe v najnovšom Slovníku súčasného slovenského jazyka. Nemalú pozornosť venovala metódam, ktoré boli použité pri spracovávaní malých slovných druhov v SSSJ, a to počítačovej podpore lexikografie – korpusovým nástrojom a programom na čistenie a zjednocovanie konceptov, a tiež vyčleňovaniu sémantických skupín malých slovných druhov.

Elizaveta R u m y a n t s e v a z Moskovskej štátnej lingvistickej univerzity v príspevku *Optimization of Russian Bilingual Dictionaries* predstavila projekt digitalizácie dvojjazyčného rusko-anglického slovníka a jeho pilotnú verziu. Tento projekt bol reakciou na nedostatok vysoko kvalitných slovníkov, ktoré by odrážali požiadavky prekladateľov. Autorka zdôraznila výhody počítačovej podpory lexikografie a možnosti, ktoré elektronický slovník poskytuje. V diskusii k príspevku sa potom rozvinula úvaha o výhodách a nevýhodách pasívnych a aktívnych elektronických slovníkov.

Rokovanie ďalej pokračovalo sekciou vedenou Karlom Palom. Autori príspevku *The Text Corpus and Dictionary Hierarchy* Natalia D a r č u k, Ľudmila A l e x e j e n k o v a a Viktor S o r o k i n z Kyjevskej národnej lingvistickej univerzity prezentovali základné princípy

tvorenia elektronickej databázy, v ktorej sa má systematicky zachytiť opis jazykových jednotiek na všetkých jazykových úrovniach. Na základe týchto princípov boli vyvinuté počítačové nástroje, ktoré zabezpečujú extrakciu špecifických lingvistických informácií z textov. Štruktúra databázy je pritom založená na tzv. modulovej lexikografickej ideológii, podľa ktorej sa systém a štruktúrne vzťahy každej úrovne jazykového systému zachytávajú v čiastkových moduloch – morfologickom, morfematickom, sémantickom, slovotvornom a syntagmatickom.

Victor Zacharov z Fakulty filológie Štátnej univerzity v Petrohrade sa v príspevku *Citation Card Files, Corpora of the Past* venoval elektronickej kartotéke a korpusu a ich úlohe v modernej lexikografii. Kartotéka Inštitútu pre lingvistické štúdie v Petrohrade (The Large Card File – LCF) obsahuje 8 miliónov excerpčných lístkov a s jej digitalizáciou sa začalo v roku 2006. LCF databáza obsahuje zoznamy slov a konkordancií**,** ako aj ich zdrojov. Databáza umožňuje získavanie štatistických údajov, ako aj porovnávanie konkordancií. Autor zdôraznil, že súčasná práca na projekte sa zameriava na vyvíjanie sémantických a rôznych ďalších filtrov, ktoré by pomohli výskumu, výberu a uchovávaniu dát a uľahčili tým prácu na slovníku.

Prednášajúci Róbert Sabo z Ústavu informatiky SAV v Bratislave v spolupráci s vedúcim oddelenia analýzy a syntézy reči tohto ústavu Pavlom Ruskom a spolupracovníkom Martinom Dzúrom z Katedry literatúry a literárnej vedy Filozofickej fakulty Univerzity Komenského v Bratislave v prezentácii s názvom *Prosody Annotation in Slovak Using Sk-ToBI* prvýkrát predstavil lingvistickej verejnosti návrh novej schémy pre intonačnú anotáciu slovenských hovorených prejavov. Inšpirovaní úspechom anotačnej schémy ToBI pre angličtinu a nemčinu sa autori rozhodli nasledovať jej hlavné princípy a definovať tak špeciálnu slovenskú verziu Sk-ToBI. Navrhovaný anotačný systém bol testovaný na databáze štúdiových nahrávok bábkového herca. Dosiahnuté výsledky intonačnej anotácie reči poukázali na ďalšie spôsoby zlepšovania anotačného systému Sk-ToBI.

Hovorený korpus ORAL 2006 bol témou Martiny Waclawičovej z Ústavu Českého národného korpusu Karlovej univerzity v Prahe. V príspevku s názvom *Spoken Corpus ORAL 2006, Information that it Provides and General Characteristics of Spoken Text* poukázala na možnosti, ktoré tento korpus poskytuje používateľom. Nová verzia predstavuje reprezentatívny, čiže vyvážený hovorený korpus obsahujúci nahrávky spontánnych hovorených prejavov. Výnimočnosťou tohto hovoreného korpusu je, že okrem špeciálnej transkripcie, pomocou ktorej sú zachytené zvukové vlastnosti textov, obsahuje informácie o sociolingvistických aspektoch textov, ktoré je možné spolu s transkripciou vyhľadávať v korpusovom manažéri Bonito.

Pilotnú verziu hovoreného korpusu slovenského jazyka predstavil v príspevku *Corpus of Spoken Slovak* Radovan Garabík z SNK JÚĽŠ SAV, na ktorom spolupracoval s Pavlom Ruskom z Ústavu informatiky SAV v Bratislave. Výsledná podoba hovoreného korpusu pozostáva z niekoľkých nahrávok s ich ortografickou transkripciou. Autor sa venoval niektorým z hlavných konceptov, návrhov a technických riešení pre zaznamenávanie a poloautomatickú anotáciu a prezentoval všeobecnú anotačnú štruktúru vhodnú pre daný typ hovoreného korpusu. Tiež zdôraznil nevyhnutnosť budovania národného hovoreného korpusu.

Poslednú popoludňajšiu sekciu vedenú Alešom Horákom otvorila príspevkom *Statistical Syllable Segmentation Precision as a Function of Training Data Quality* Daniela Majchráková z SNK JÚĽŠ SAV, na ktorom spolupracovala s Jozefom Ivaneckým z European Media

Laboratory v Heidelbergu. Príspevok bol zameraný na zdokonaľovanie štatistického prístupu pri určovaní slabičných hraníc. Cieľom bolo dosiahnutie čo najlepšej štatistickej úspešnosti, pričom sa uvažovalo o kvalite a kvantite trénovacích a testovacích dát použitých pre slabičnú segmentáciu. Štatistické výsledky viedli k záveru, že vzhľadom na limitované tréningové dáta výber dát je len čiastočne dôležitý.

Adam Przepiórkowski z Inštitútu informatiky Poľskej akadémie vied v príspevku *Automatic Valence Acquisition in Polish* prezentoval výsledky projektu *Automatic extraction of linguistic knowledge from a large corpus of Polish* (Automatická extrakcia lingvistickej informácie z veľkého poľského korpusu). Venoval sa opisu metodológie automatického určovania valencie v morfosyntakticky anotovanom korpuse, vysvetlil jednoduché gramatické pravidlá, na základe ktorých boli vo vetách identifikované potenciálne argumenty slovesa, a tiež štatistické techniky, ktoré sa aktuálne použili na zredukovanie chýb vo výslednom valenčnom slovníku.

Problematike kolokácií sa venoval Peter Ďurčo z Univerzity sv. Cyrila a Metoda v Trnave v príspevku *Collocations in Slovak (Based on the Slovak National Corpus)* a zároveň dal do pozornosti pripravovaný projekt Konfrontačného nemecko-slovenského slovníka kolokácií. Slovník bude obsahovať viacslovné ustálené spojenia, frazémy a typické spojenia, pri ktorých bude uvedená akj kvantifikácia výskytu, aj frekvencia. Materiálovým základom slovenskej časti slovníka budú dáta Slovenského národného korpusu, pri analýze spojení sa budú využívať nástroje a metódy korpusovej lingvistiky. Jednotlivé kolokácie budú priraďované ku kolokačným vzorcom, pričom slovník bude zachytávať kombinatorický potenciál substantív, slovies**,** adjektív a adverbií.

Posledný deň konferencie v sekcii vedenej Radovanom Garabíkom zahájila svojím príspevkom *Collocations in Russian: Analysis of Association Measures* Mária Chochlova z Fakulty filológie Štátnej univerzity v Petrohrade. Autorka sa pokúsila o krátky prieskum rozličných prístupov k problematike kolokácií (britský kontextualizmus, lexikografický prístup, teória „Meaning-Text“) na príkladoch niektorých kolokácií vybraných z korpusu ruských textov. Tvrdí, že aplikácia korpusových metód určovania lexikálnej kolokability slov môže viesť k vytvoreniu slovníka kolokácií. Tiež navrhuje, aby sa informácie o kolokabilite slov založené na štatistických hodnotách odrazili v súčasných ruských slovníkoch.

Korpusovej analýze spájateľnosti slov v ruštine sa venovala Olga Mitrofanova z Fakulty filológie Štátnej univerzity v Petrohrade v príspevku *Corpus Analysis of Selectional Preferences in Russian*, ktorý predniesla za spolupracovníčky Viktoria Bielik a Veru Kadinu. Vo svojom projekte sa zamerali na výskum verbálnych a nominálnych fráz v korpuse ruských textov. Analyzovali spoluvýskyty jednotlivých lexikálnych jednotiek v rámci bigramu na základe hodnôt MI-score. Na zhodnotenie štatistických mier MI-score boli vytvorené špeciálne pravidlá Optimality Theory rules. Výsledkom bolo vytvorenie zoznamu najfrekventovanejších slov a ich kolokačných matíc – zoznamov slov, s ktorými daná lexikálna jednotka vytvára kolokáciu.

Aj nasledujúci príspevok *Automatic Word Clustering in Russian Texts Based on Latent Semantic Analysis* prezentovala Olga Mitrofanova, tentoraz za kolegov Polinu Paniševu a Viačeslava Savitskeho z Fakulty filológie Štátnej univerzity v Petrohrade. Autorka sa v ňom zaoberá vyvinutím a aplikáciou nástroja na automatické zhlukovanie slov (AWC), ktorý slúži na spracovanie ruských textov rôznych typov. Vytvorenie AWC nástroja vyžadovalo

počítačovú implementáciu LSA (Latent Semantic Analysis) v kombinácii s klastrovacím algoritmom. Pre potreby tohto projektu bol tiež vyvinutý softvér na báze jazyka Python. Hlavné postupy tvorené nástrojom AWC sú segmentácia vstupných textov, kontextová analýza či vytváranie kolokačných matíc.

Poslednú sekciu konferencie vedenú Jozefom Ivaneckým otvorila príspevkom *Systemic and Functional Features of the Ukrainian Nouns Category of Number* Tatyana B o b k o v a z Kyjevskej národnej lingvistickej univerzity. Autorka sa zaoberala systémovými a funkčnými vlastnosťami kategórie čísla ukrajinských podstatných mien. Výskum uskutočnila na oficiálnych textoch NATO. Z nich bol vytvorený mikroslovník obsahujúci 20 tisíc substantív a ich slovných tvarov, ktoré sa následne štatisticky vyhodnocovali. V analyzovanom materiáli sa najčastejšie vyskytovali substantíva, pre ktoré je charakteristická singulárová aj plurálová podoba, najnižšiu frekvenciu mali substantíva pluralium tantum.

Problematika automatickej identifikácie prirodzených jazykov bola témou príspevku Petra V o j t e k a a Márie B i e l i k o v e j z Fakulty informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave: *Comparing Natural Language Identification Methods Based on Markov Processes*. Porovnali dva prístupy identifikácie jazykov založené na Markovovej metóde. Obidve metódy spracovávajú vstupný text na úrovni znaku. Dané prístupy sa experimentálne overovali na Multilingual Reuters Corpus, ktorý obsahuje texty rozličných európskych jazykov. Použitím týchto metód boli jednotlivé jazyky identifikované s 99,75 %-nou úspešnosťou.

Predposledným príspevkom konferencie bola prednáška s názvom *Hyperlemma: a Concept Emerging from Lemmatizing Diachronic Corpora* autora Karla K u č e r u z Ústavu Českého národného korpusu Karlovej univerzity v Prahe. Pri lematizácii diachrónneho korpusu, ktorý zahŕňa texty z celej histórie jazyka, bolo potrebné adaptovať širší pojem pre lemmu a uchopiť tak rozmanitosť morfologických, fonologických a ortografických foriem. Na základe týchto potrieb bol do lematizátora pre diachrónnu časť ČNK implementovaný koncept hyperlemmy. Autor sa venoval definícii pojmu hyperlemma, ktorý zahŕňa nielen všetky ohybné tvary slova, ale na rozdiel od lemmy aj historické a dialektologické fonologické varianty, ako aj moderné pravopisné varianty.

Súčasný stav terminológie a nové aktivity v tejto oblasti predstavila Jana L e v i c k á z SNK JÚĽŠ SAV v príspevku *Building Slovak Terminology Database: Definition and Defining Context Evaluation*. Predstavila rozbiehajúci sa projekt Slovenskej terminologickej databázy (STD) realizovaný oddelením SNK JÚĽŠ SAV, ktorý vzišiel z potrieb odbornej komunikácie na Slovensku. STD je sprístupnená verejnosti zatiaľ skúšobne na webovej stránke SNK. Jednotlivé terminologické vstupy sú dopĺňané kontextovými definíciami dostupnými na internete. V budúcnosti sa počíta s vytvorením podkorpusu odborných textov príslušných oblastí, z ktorého by terminologická databáza čerpala vlastné definície terminologických hesiel.

Záverečnou bodkou konferencie bol príhovor hlavnej organizátorky Jany L e v i c k e j, ktorá sa všetkým prítomným poďakovala za účasť a zároveň ich pozvala na nasledujúce SLOVKO 2009, ktoré bude venované korpusovej lingvistike.

Adriana Oravcová – Daniela Majchráková